CAPITOLO 2

L’approccio metodologico

Luigi Mastronardi, Gianluca Monturano, Luca Romagnoli, Mara Vasile, Mariella Zingaro[1]

2.1 Le fasi del percorso metodologico

Il percorso metodologico qui sperimentato è articolato in tre stadi (Figura 2.1):

1. Individuazione di gruppi di comuni omogenei per grado di vulnerabilità;
2. Individuazione delle principali motivazioni alla base della costituzione delle cooperative di comunità;
3. Studio di scenario per definire il ruolo potenziale della cooperativa a supporto delle motivazioni comunitarie.

Nella Fase 1., vengono individuati gli elementi che concorrono a formare il tessuto sociale e produttivo regionale e a far emergere le “risorse latenti” presenti sul territorio per uno sviluppo locale endogeno. La comprensione delle strutture e dei processi è alla base dello studio.

[1] A Luigi Mastronardi insieme a Luca Romagnoli si deve la redazione del paragrafo 2.1, a Luca Romagnoli quella del paragrafo 2.2, a Mariella Zingaro dei paragrafi 2.3 e 2.4, a Mara Vasile il paragrafo 2.5 e a Gianluca Monturano il paragrafo 2.6.

Luigi Mastronardi, University of Molise, Italy, luigi.mastronardi@unimol.it, 0000-0001-6012-2964
Gianluca Monturano, University of Molise, Italy, monturano@hotmail.it
Luca Romagnoli, University of Molise, Italy, luca.romagnoli@unimol.it, 0000-0003-3243-1561
Mara Vasile, University of Molise, Italy, maravasile@live.it
Mariella Zingaro, University of Molise, Italy, mariellazingaro@outlook.it

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Luigi Mastronardi, Gianluca Monturano, Luca Romagnoli, Mara Vasile, Mariella Zingaro, *L’approccio metodologico*, pp. 59-90, © 2020 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-168-6.03, in Luigi Mastronardi, Luca Romagnoli (edited by), *Metodologie, percorsi operativi e strumenti per lo sviluppo delle cooperative di comunità nelle aree interne italiane*, © 2020 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5919 (online), ISBN 978-88-5518-168-6 (PDF), DOI 10.36253/978-88-5518-168-6

Figura 2.1. Fasi del percorso metodologico-operativo

I metodi di analisi permettono di selezionare le variabili strutturali rispetto all'interrelazione tra componente sociale, economica, istituzionale e ambientale; in questo modo, è possibile comprendere le interazioni che sono alla base dei processi esistenti e inespressi, ossia, la capacità portante di un determinato territorio e la risposta possibile alle perturbazioni.

Questo schema metodologico permette di procedere alla classificazione dei comuni di una determinata regione, al fine di tracciare una geografia del grado di vulnerabilità/resilienza funzionale e delineare gli scenari in cui procedere alla costituzione delle cooperative di comunità.

Nel dettaglio tale fase è orientata a: 1) individuare gli strumenti metodologici per classificare il territorio; 2) identificare e costruire indicatori di sintesi che declinino il paradigma interpretativo individuato per conoscere e valutare l'entità del patrimonio economico, sociale e ambientale presente, anche attraverso macro aggregazioni territoriali e indici di sintesi.

Nella Fase 2., le attività di ricerca sono dirette a comprendere le esigenze delle comunità locali in quanto portatrici di bisogni, con particolare riferimento al settore agricolo e ai servizi di cura e gestione del territorio e delle risorse ambientali.

In sostanza, le analisi qui condotte sono finalizzate a far emergere le principali motivazioni comunitarie che possono essere alla base della costituzione delle cooperative di comunità, e in relazione alle quali queste possono offrire risposte adeguate alla risoluzione delle criticità territoriali. A tal fine, è condotta un'analisi quali/quantitativa per individuare i fabbisogni e di conseguenza le condizioni utili a definire il ruolo e gli ambiti di attività delle cooperative di comunità nei diversi territori di riferimento.

I fabbisogni delle comunità locali vengono rilevati mediante indagine diretta condotta con questionari strutturati, somministrati ai testimoni privilegiati legati alle singole realtà locali.

L'indagine è svolta su alcune aree pilota individuate sulla base dei risultati ottenuti nella prima fase dello studio e in considerazione degli ambiti operativi di alcuni istituti finanziari presenti a livello locale.

Sempre in questa fase, lo sforzo metodologico è finalizzato a comprendere l'importanza e il contributo delle cooperative di comunità che promuovono forme diverse di produzione e consumo in relazione ai fabbisogni del territorio, ovvero l'offerta di moderni servizi di welfare, e in generale di beni e servizi di interesse collettivo, la valorizzazione dei patrimoni pubblici e privati inutilizzati, la gestione dei beni ambientali e culturali e altre attività necessarie per innervare processi di sviluppo e per il miglioramento delle condizioni di benessere delle popolazioni locali. Pertanto, sono individuati i fattori che possono caratterizzare le CdC per chiarirne le peculiarità e gli aspetti funzionali e organizzativi che le contraddistinguono nei processi di sviluppo su base comunitaria, e per valorizzare il potenziale di imprenditoria "nascosta" nei territori.

Si tratta in sostanza di definire i principali requisiti che caratterizzano le cooperative di comunità sia rispetto ai modelli cooperativi tradizionali, sia per giustificare la loro attività chiaramente diretta al benessere delle comunità locali.

A valle di questa analisi, viene definito un modello di cooperativa di comunità in relazione alle diverse configurazioni che essa può assumere per quanto riguarda il soddisfacimento dei fabbisogni delle comunità e di conseguenza alle attività da essa esercitata nel contesto territoriale di riferimento.

Tali attività di ricerca permettono di acquisire informazioni da utilizzare nella fase successiva dello studio, per definire la costituzione di adeguati stereotipi cooperativi con riferimento alla struttura, alle funzioni, ai servizi e all'organizzazione degli stessi.

Nella Fase 3., è valutata la "fattibilità" degli elementi di definizione del modello di cooperativa di comunità, in considerazione del fatto che vi possono essere diverse tipologie cooperative, in relazione ai bisogni, agli interessi, alla dimensione della comunità locale e di conseguenza al tipo di bene e/o servizio prodotto. È importante dunque comprendere sempre in questa fase la realizzabilità della CdC per quanto riguarda la dimensione tecnica, economico-finanziaria, organizzativa, giuridica.

Le analisi sono inoltre orientate all'individuazione degli strumenti di finanziamento utili a favorire la costituzione e lo sviluppo delle cooperative di comunità.

2.2 L'analisi statistica multivariata

In questa fase, sono identificate le variabili e costruiti indicatori di sintesi per conoscere il livello di vulnerabilità sociale, economica, istituzionale e ambientale del territorio oggetto di studio, attraverso macro aggregazioni territoriali e indici di sintesi.

Lo studio si può basare su un set di indicatori di "forza" o al contrario di "debolezza" (vedi un'esemplificazione in Tabella 2.1) che riguardano fenomeni relativi ai profili socio-demografico, economico, ambientale e istituzionale.

Gli indicatori sono facilmente reperibili, e possono essere estrapolati prevalentemente dai censimenti Istat della popolazione, dell'agricoltura e dell'industria, integrati da altre fonti di pubblica consultazione (ISPRA, archivi degli Enti pubblici). Gli indicatori demografici e occupazionali sono quelli più utilizzati negli studi sulle dinamiche della popolazione e il mercato del lavoro. La vulnerabilità nella struttura economica, nel sistema dell'istruzione e socio-assistenziale e sanitaria è rappresentata dalle variabili relative ai comparti dell'economia fondamentale. Essa comprende le attività territorializzate, cioè legate a contesti locali, i cui prodotti vengono usati, tendenzialmente, da tutti i cittadini, a prescindere dal reddito di cui dispongono (Barbera *et al.*, 2016). L'economia fondamentale comprende comparti come la produzione e la distribuzione di cibo, i servizi sanitari e di cura, l'istruzione, i trasporti, la distribuzione di energia, di acqua e di gas, le telecomunicazioni, la raccolta e il trattamento dei rifiuti.

L'impianto metodologico avvale di due note tecniche statistiche multivariate, quali l'*Analisi delle Componenti Principali* (ACP) e l'*Analisi dei Gruppi* (Cluster Analysis, CA), per:

1) Individuare aree omogenee sul territorio secondo il grado di vulnerabilità sociale, economica, istituzionale e ambientale;
2) Percepire i fabbisogni delle comunità locali;
3) Percepire le "risorse latenti" presenti nel territorio;
4) Delineare le linee d'intervento delle cooperative di comunità;
5) Individuare ambiti territoriali ideali per realizzare economie di scala.

L'ACP (Fabbris, 2011) è una metodologia statistica multivariata che, partendo da una matrice dei dati di dimensioni $(n \times p)$ (dove n rappresenta il numero delle unità statistiche e p il numero delle variabili) con variabili tutte quantitative, consente di sostituire alle variabili originali (tra loro correlate) un nuovo insieme di variabili, chiamate componenti principali (CP), che godono delle seguenti proprietà:

1. Sono tra loro incorrelate (ortogonali);
2. Sono elencate in ordine decrescente della loro varianza (Zani e Cerioli, 2007).

La logica sottostante il metodo è che, in una analisi multivariata, una variabile è tanto più rilevante, quanto più è elevata la sua variabilità (misurata dalla varianza), poiché ciò significa che le unità statistiche osservate sono fra di loro molto differenziate in termini della variabile considerata. Si richiede, inoltre, che le nuove variabili (le CP) siano fra loro incorrelate, perché in questo modo ciascuna di esse potrà fornire il massimo delle informazioni possibili: è noto, infatti, che quanto più due variabili sono correlate, tanto più esprimono la stessa informazione, presentando di conseguenza informazioni ridondanti. Proprio per quest'ultimo motivo, maggiore è la correlazione tra le variabili, minore sarà

il numero di componenti principali che verranno estratte. Le proprietà fondamentali delle CP estratte sono:

- La v-esima componente principale di p variabili, espresse in termini di scostamenti dalla media, è data dalla formula: $y_v = \tilde{X} a_v$, per v=1, … , $k \leq p$, dove $\tilde{X} = [x_{is} - \bar{x}_s]$ è la matrice (n×p) degli scarti dei valori osservati dalle proprie medie di variabile, ossia di colonna, e a_v è l'autovettore associato al v-esimo autovalore λ_v (in ordine decrescente) della matrice di covarianza: ciò significa che ogni CP è una combinazione lineare delle variabili originarie;

Tabella 2.1. *Indicatori per la classificazione del territorio regionale*

Cod.	**Variabili demografico-sociali**
D1	Popolazione residente
D2	Densità demografica
D4	Incidenza popolazione residente con meno di 6 anni
D5	Incidenza popolazione residente di 75 anni e più
D8	Indice di vecchiaia
D10	Incidenza di anziani soli
D12	Potenzialità d'uso abitativo nei centri abitati
D14	Consistenza delle abitazioni storiche occupate
D15	Incidenza di adulti con diploma o laurea
D16	Incidenza di giovani con istruzione universitaria
D18	Incidenza di adulti con lic. media
D21	Incidenza giovani 15-29 anni che non studiano e non lavorano
D22	Tasso di disoccupazione masch.
D23	Tasso di disoccupazione femm.
D24	Tasso di disoccupazione
D25	Tasso di disoccupazione giovan.
D26	Incidenza dell'occupazione nel settore agricolo
D27	Incidenza dell'occupazione nel settore industriale
D28	Incidenza dell'occupazione nel settore terziario extracommercio
D33	Mobilità occupazionale
D34	Mobilità studentesca
D40	Incidenza delle famiglie con potenziale disagio economico
D41	Incidenza di famiglie in disagio di assistenza
	Variabili economiche
E1	Reddito pro-capite
E2	U.L. commercio al dettaglio in esercizi non specializzati

E3	U.L commercio al dettaglio di prodotti alimentari in esercizi specializzati
E4	U.L. commercio al dettaglio di carburante in esercizi specializzati
E5	U.L. trasporti di passeggeri
E6	U.L. servizi postali e di corriere
E7	U.L. ristoranti e attività di ristorazione mobile
E8	U.L. bar e altri esercizi simili senza cucina
E10	U.L intermediazione monetaria
E11	U.L istruzione prescolastica
E12	U.L istruzione primaria
E13	U.L. istruzione secondaria
E14	U.L. servizi degli studi medici e odontoiatrici
E15	U.L. altri servizi di assist. sanitaria
E18	U.L. farmacie
	Variabili ambientali
A3	SAT (Sup. agricola totale)/ST
A5	Superficie biologica
A6	Superfici DOP/IGP
A7	Superfici usi civici
A8	SAT non utilizzata
A9	Superfici tratturali (mq)
A10	Superficie in dissesto (Pop. esposta a frane)
A11	N. centrali elettriche
A12	Superficie forestale
	Variabili istituzionali
I5	Associazioni non profit

- Ogni autovalore λ_v è uguale alla varianza della corrispondente v-esima componente principale.
- Il coefficiente di correlazione lineare tra la v-esima componente principale e la s-esima variabile è: $r(Y_v, X_s) = r_{vs} = \frac{a_{vs}\sqrt{\lambda_v}}{\sqrt{var(X_s)}}$.

Se si lavora su variabili standardizzate, invece, la matrice delle osservazioni è la $Z = \frac{x_{is} - \bar{x}_s}{\sigma_s}$, dove $\sigma_s = \sqrt{var(X_s)}$ è lo scarto quadratico medio della s-sima variabile, e vale quanto segue:

- La v-esima componente principale di p variabili standardizzate è data dalla combinazione lineare: $y_v = Z_{av}$, per $v = 1, \ldots, k \leq p$ in cui a_v è l'autovettore associato al v-esimo autovalore λ_v (in ordine decrescente) della matrice di correlazione.

- La somma degli autovalori è uguale a p, e la quota di varianza totale spiegata dalla v-esima componente principale è uguale a $\frac{\lambda_v}{p}$.
- Il coefficiente di correlazione tra la v-esima componente principale e la s-esima variabile è: $r_{vs} = a_{vs}\sqrt{\lambda_v}$.
- La quota di varianza della s-esima variabile spiegata dalle prime k componenti principali è uguale a: $\sum_{v=1}^{k} r_{vs}^2$ per $s = 1, \dots, p$.

In termini applicativi, lo scopo di una ACP è quello di ottenere, a partire da un consistente numero di variabili originarie, un numero (piccolo) di variabili "latenti" o "artificiali", ciascuna delle quali raccolga in sé la più elevata quota possibile della varianza complessiva, e che riesca a spiegare, da sola, un aspetto importante del fenomeno osservato: le variabili originarie che contribuiscono in misura fondamentale alla determinazione di ogni CP vengono individuate attraverso i coefficienti di correlazione fra le variabili e ciascuna CP. È chiaro che l'analisi viene svolta con l'utilizzo di software statistici; grazie a questi programmi, una volta inserito il database, si otterrà una descrizione accurata del numero di componenti estratte e di alcune informazioni indispensabili, quali la comunalità[2] e la varianza totale spiegata.

Il passo metodologico successivo, cioè la Cluster analysis (Kaufman e Rousseeuw, 2005) viene implementato considerando come input proprio le CP evidenziate nel primo step. Scopo fondamentale della CA è quello di individuare le unità amministrative territoriali più simili fra loro rispetto alle variabili considerate o, il che è (approssimativamente) lo stesso, rispetto alle CP estratte. Il concetto di "distanza" si riferisce a quello matematico-statistico di distanza fra 2 unità statistiche, che vengono misurate da un insieme di variabili quantitative. Formalmente:

La distanza tra due unità statistiche $x, y \in R^p$ è definibile come una funzione $d(x,y)$ che gode delle proprietà di:

1. *non negatività:* $d(x,y) \geq 0 \quad \forall\, x, y \in R^p$
2. *identità:* $d(x,y) = 0 \quad \leftrightarrow \quad x = y$
3. *simmetria:* $d(x,y) = d(y,x) \quad \forall\, x, y \in R^p$
4. *disuguaglianza triangolare:* $d(x,y) \leq d(x,z) + d(y,z)$, con: $x, y, z \in R^p$

Esistono diversi tipi di distanze, tutte riconducibili alla distanza di Minkowsky di k-esimo ordine tra le unità i e j, descritta dalla seguente espressione:

$$ {}_k d_{ij} = \left[\sum_{s=1}^{p} \left| x_{is} - x_{js} \right|^k \right]^{1/k} \qquad k \geq 1 $$

Le distanze più note sono la *distanza euclidea* (k=2) e quella di Manhattan (k=1). Schematicamente, una CA consta di cinque passaggi fondamentali:

[2] Quote di varianza di ogni variabile (standardizzata) spiegate dalle prime componenti principali estratte.

1. Scelta delle variabili d'interesse: in prima battuta il ricercatore è chiamato a selezionare logicamente le variabili da sottoporre ad analisi statistica. Quando le variabili sono molto numerose, è possibile ricorrere, come nel nostro caso, all'analisi delle componenti principali, in maniera tale da ottenere un numero ridotto di variabili sulle quali lavorare, senza eccessiva perdita di informazione.
2. Scelta della distanza o dell'indice di similarità: come descritto in precedenza, esistono vari tipi di distanza. Il ricercatore è tenuto a effettuare una scelta soggettiva, che comunque può incidere sul risultato finale.
3. Scelta del metodo di formazione dei gruppi: l'obiettivo della CA è quello di riuscire a formare gruppi di unità statistiche omogenei al loro interno, ma eterogenei tra loro (massima varianza fra i gruppi, o cluster, e minima varianza all'interno dei cluster). Esistono due metodi di formazione dei gruppi: a) *gerarchici* e b) *non gerarchici*.

a. I metodi gerarchici sono quelli in cui ogni unità osservata costituisce all'inizio un cluster a sé stante – ci sono n cluster di 1 unità ciascuno -. I due "cluster" (unità) più vicini (cioè quelli che presentano la minore distanza all'interno della cosiddetta "matrice delle distanze" calcolata al passo 2) vengono uniti, e ciò viene fatto ripetutamente fino a quando tutte le unità considerate sono in un unico cluster. Di conseguenza, l'output finale di questi metodi non è una singola partizione delle n unità, bensì una serie di partizioni, che vengono rappresentate graficamente per mezzo di un *dendrogramma*, che contiene i livelli di distanza sull'asse verticale, e le singole unità su quello orizzontale. La linea orizzontale che unisce due o più "rami" evidenzia il livello di distanza a cui due cluster si uniscono. I metodi gerarchici differiscono, in particolare, nel modo in cui le distanze vengono ricalcolate fra il nuovo cluster che è appena formato e i cluster rimanenti dopo la k-sima fusione.
Indicando con C_a e C_b due generici cluster composti, rispettivamente, da n_a e n_b unità; con i e l due singole unità (con $i \in C_a$ e $l \in C_b$); e con $d(C_a,C_b)$ la distanza fra i cluster C_a e C_b, i più utilizzati metodi gerarchici sono i seguenti:

- *Legame singolo*: $d(C_a, C_b) = \min_{i \in C_a; l \in C_b} d_{il}$;
- *Legame completo*: $d(C_a, C_b) = \max_{i \in C_a; l \in C_b} d_{il}$;
- *Legame medio*: $d(C_a, C_b) = \frac{1}{n_a n_b} \sum_{i \in C_a} \sum_{j \in C_b} d_{il}$;
- *Metodo del centroide*: $d(C_a, C_b) = d(\bar{\mathbf{z}}_a, \bar{\mathbf{z}}_b)$, dove $\bar{\mathbf{z}}_a$ e $\bar{\mathbf{z}}_b$ sono i centroidi dei due cluster – cioè i vettori dei valori medi delle p variables nei cluster C_a e C_b:

$$\bar{z}_{j;(cl)} = \frac{1}{n_{cl}} \sum_{i \in C_{cl}} z_{ij} \text{ , per e } cl = a,b \text{ e } j = 1, \dots, p$$

- *Metodo di Ward*. Calcoliamo le quantità:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(z_{ij} - \bar{z}_j\right)^2$$

dove $\bar{z}_j$ è la media della j-sima variabile nell'intero insieme delle osservazioni – essa è pari a 0 quando si tratta, come solitamente accade, di variabili standar-

dizzate, cioè variabili a media 0 e varianza 1 –; data una partizione in g cluster, la devianza totale, T, può essere decomposta in:

$$W = \sum_{k=1}^{g} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left(z_{ij} - \bar{z}_{j;(k)}\right)^2$$

dove $\bar{z}_{j;(k)}$ è la media della j-sima variabile nel cluster k; e

$$B = \sum_{k=1}^{g} \sum_{j=1}^{p} \left(\bar{z}_{j;(k)} - \bar{z}_j\right)^2$$

in base alla ben nota relazione: $T = W + B$. Il metodo di Ward si basa sul fatto che, passando dalla partizione in $k + 1$ cluster a quella in k cluster, W (devianza complessiva entro i gruppi) tende ad aumentare (minore omogeneità nel nuovo cluster che si è creato, per via dell'aggiunta di una unità), mentre naturalmente B (devianza complessiva fra i gruppi) diminuisce: a ogni passo della procedura di Ward, i cluster che si uniscono saranno i due per i quali l'aumento nella quantità W sarà minimo.

b. I metodi *non gerarchici*, invece, forniscono direttamente un'unica partizione delle n unità in un numero di gruppi fissato a priori dal ricercatore. Il meccanismo secondo cui allocare le unità dipende da una funzione obiettivo solitamente espressa in termini di scomposizione della devianza totale. In questo modo si cerca di ottenere una partizione che abbia il requisito della massima coesione nei gruppi. Tali metodi hanno il vantaggio di poter essere applicati a un numero molto elevato di unità in quanto non richiedono il calcolo della matrice delle distanze. Inoltre, l'assegnazione a un gruppo non è definitiva (come nei metodi gerarchici) in quanto il ricalcolo della funzione obiettivo può comportare lo spostamento dell'unità da un cluster a un altro se questo assicura una maggiore coesione interna. Lo svantaggio consiste ovviamente nella necessità di individuare a priori il numero di gruppi da assegnare alla partizione.

4. Individuazione del numero ottimo di gruppi: una volta costruito il dendrogramma associato a un'analisi (gerarchica), il ricercatore deve decidere a che livello sezionare tale grafico, ottenendo così un numero di gruppi che soddisfi le esigenze del caso. Una regola pratica potrebbe essere quella di prendere in considerazione l'aumento relativo nella distanza di fusione di due cluster, δ_k, per $k = n-1, n-2, \dots, 1$: data la partizione in $k + 1$ cluster (col proprio livello di distanza d_{k+1}), e la successiva partizione in k clusters (col proprio livello di distanza d_k, per la quale vale sicuramente la relazione: $d_k \geq d_{k+1}$), è possibile calcolare il valore relativo: $\delta_k = (d_k - d_{k+1})/d_{k+1}$, e scegliere il numero di cluster per il quale δ_k è massimo.
   Nelle applicazioni, è pratica comune ripetere l'analisi per differenti numeri di cluster, e quindi calcolare la funzione obiettivo:

$$R_{(k)}^2 = \frac{B_{(k)}}{T}$$

   dove $B_{(k)}$ è la devianza fra i gruppi (si veda il punto precedente) relativa alla partizione in k cluster, e T, come prima, è la devianza totale.
   Riportando su un grafico il numero di cluster, k (sull'asse orizzontale) e i valori $R_{(k)}^2$ (sull'asse verticale), una buona scelta per il numero di cluster sarà il

valore k per cui il grafico presenta un "gomito" (ossia una repentina riduzione nella pendenza): ciò significa che si ferma la scelta al numero di cluster per i quali un aumento da k a $k + 1$ porta a un aumento "piccolo" del rapporto devianza fra i gruppi/devianza totale.

5. Un controllo di robustezza della partizione finale prescelta può essere condotto con l'utilizzo del noto indice di Rand, che permette di calcolare il grado di "concordanza" di 2 partizioni, determinate sulla base dell'impiego di diverse matrici di distanze e/o di diversi algoritmi di aggregazione. In particolare, date due partizioni, P e P^*, rispettivamente di g e g^* cluster, ottenute con l'applicazione di diverse metodiche di clustering sulle stesse n unità, l'indice di Rand viene calcolato come:

$$R_{P,P*} = 1 - \frac{\sum_{i=1}^{g} n_{i0}^2 + \sum_{j=1}^{g^*} n_{0j}^2 - 2\sum_{i=1}^{g}\sum_{j=1}^{g^*} n_{ij}^2}{n(n-1)}$$

dove n_{i0} è il numero di unità appartenenti all'i-esimo cluster nella partizione P; n_{0j} è il numero di unità appartenenti al j-esimo cluster nella partizione P^*; e n_{ij} è il numero di unità appartenenti congiuntamente al cluster i nella partizione P, e al cluster j nella partizione P^*. L'indice varia da 0 a 1: vale 0 se ciascuna unità appartiene a cluster diversi nelle due partizioni; vale 1 se le partizioni sono identiche. È chiaro che una misura di robustezza valida sarà quella derivante dal confronto di due partizioni non troppo differenti: molto spesso si calcola l'indice di Rand fra partizioni che, a parità di numero di cluster ($g = g$), sono state determinate sulla base delle sole matrici di distanza differenti o, ancor più di frequente, attraverso due diversi algoritmi di aggregazione. L'ultima fase è quella relativa alla caratterizzazione dei cluster risultanti dall'applicazione della procedura seguita nelle fasi 1-4. In particolare, ciò che viene solitamente fatto è riportare in una tabella le medie delle variabili originarie per ciascun cluster, allo scopo di evidenziare le principali caratteristiche dei gruppi in relazione ai macro ambiti investigati.

2.3 La progettazione e somministrazione dei questionari

Il procedimento metodologico finora delineato ha consentito, mediante un approccio quantitativo, di comprendere il territorio regionale nella sua completezza e di poter definire le aree specifiche oggetto di studio. Sulla spinta dei risultati quantitativi raggiunti e sull'esigenza di conoscere nel dettaglio il territorio, per l'eventuale avvio di cooperative di comunità, appare necessario che l'indagine ampli il proprio asset metodologico. La ricerca diviene a carattere qualitativo con la messa a punto di un'indagine diretta sul campo per conoscere le opinioni di quei soggetti definibili portatori di interesse delle rispettive zone individuate. Gli intervistati sono stati individuati in relazione alla loro peculiare conoscenza e competenza del territorio e allo status e al ruolo che ricoprono.

La metodologia di ricerca adottata vuole far emergere una visione della regione più precisa e accurata mediante l'implementazione di un questionario semi-strutturato da somministrare ai testimoni privilegiati tramite delle interviste dirette.

Lo sviluppo imprenditoriale delle cooperative di comunità necessita del coinvolgimento di specifici soggetti territoriali e della stipula con essi di relazioni essenziali per beneficiare di quelle risorse e infrastrutture a cui la stessa dovrà appoggiarsi (la pubblica amministrazione, le parrocchie, le associazioni territoriali, le proloco, gli imprenditori locali), principalmente nella sua prima fase di sviluppo. Per le interviste sono stati individuati, quali testimoni privilegiati: i sindaci di ogni comune selezionato, un parroco su ogni "macro-zona", il direttore della C.I.A. – Confederazione Italiana Agricoltori del Molise – e alcuni stakeholders legati al Terzo Settore. La progettazione del questionario, lo strumento fondamentale per la raccolta delle informazioni rilevanti per la ricerca, si costruisce in relazione a una precisa metodica. Nel processo metodologico la fase di formulazione delle domande riveste un'importanza fondamentale, in quanto a seconda di come esse vengono poste potrebbero influenzare, in parte o del tutto, la risposta. Le domande del questionario, per ottenere dei risultati effettivamente efficaci e pertinenti allo studio, sono state realizzate tenendo in considerazione alcuni aspetti salienti: le finalità conoscitive della ricerca (identificare i presupposti per la nascita delle cooperative di comunità in Molise e il ruolo che potrebbero ricoprire nel processo di sviluppo locale); le tematiche da indagare (le problematiche e le risorse delle Aree Interne molisane); le variabili da convalidare in relazione ai risultati quantitativi precedentemente ottenuti (vulnerabilità sociali ed economiche; elementi paesaggistici-naturalistici di rilevante interesse comunitario); i destinatari/rispondenti del questionario; l'ordine di presentazione delle domande (come filo logico da seguire nella concretizzazione di un progetto imprenditoriale, dalla lettura e analisi del territorio, alla comprensione delle problematiche emergenti e delle risorse su cui investire). Un'ampia serie di domande è stata vagliata e, infine, ridotta, affinché ciascun quesito fosse strettamente collegato agli obiettivi specifici della ricerca, alle possibili risposte che gli intervistati avrebbero potuto fornire e in prospettiva della metodologia di analisi statistica (modelli di analisi testuale) da applicare alle informazioni ottenute. Il questionario, pertanto, definiti quali obiettivi fossero essenziali, quali sarebbero stati superflui e quali sarebbero state in linea massima le risposte attese, consta di sette domande aperte. Le domande che costituiscono il questionario sono state ponderate e strutturate anche per evitare che i risultati potessero essere alterati e inattendibili, poiché la modalità aperta della risposta (essenziale per richiedere opinioni personali e approfondite) può sempre presentare un riscontro inaspettato (non prevede una gamma di risposte predeterminate). Il vantaggio della risposta aperta, in ogni caso, offre agli intervistati l'opportunità di parlare ed esprimersi liberamente. Sebbene sia presente un questionario, in funzione di griglia o traccia fissa per l'indagine diretta, identico per ogni testimone privilegiato,

la conduzione dell'intervista inevitabilmente è stata ampliata da domande suppletive in relazione alle risposte fornite dall'intervistato. Infatti, l'intervistatore ha trattato alcuni argomenti, nati spontaneamente durante l'intervista, in maniera più approfondita, ogni qual volta l'abbia ritenuto necessario, per ottenere risposte maggiormente precise o per agevolare la comprensione del quesito al rispondente. Il questionario costituisce uno schema di fondo che tocca i temi essenziali della ricerca, che inevitabilmente devono essere indagati, ma entro il quale vi è sempre una libertà di linguaggio, terminologia e atteggiamento nel trattare ogni argomento. La forma del questionario è rilevante tanto quanto il suo contenuto, per cui le domande sono state impostate in modo che seguano una continuità logica, purché il passaggio da un quesito all'altro sia chiaro e distinto. La successione delle domande segue un processo che può essere definito deduttivo, si passa da domande generali legate alla lettura attenta del Comune in questione, a domande più particolari legate invece alle cooperative di comunità, al fine di dare la possibilità al rispondente di focalizzare l'attenzione in maniera graduale sul tema proposto. Il questionario focalizza, dapprima, l'attenzione sulle criticità e sui bisogni presenti in un territorio, che potrebbero essere ridotti dalla presenza di una cooperativa di comunità, per poi comprendere quali possano essere i beni patrimoniali di cui la cooperativa di comunità potrebbe disporre e quali i servizi collettivi che potrebbe svolgere e garantire all'interno del comune di appartenenza e in quelli limitrofi. Il questionario è stato considerato definitivo solo a seguito di un test preventivo, realizzato mediante un'intervista pilota. Il questionario è stato somministrato a un sindaco di un comune non oggetto della ricerca. L'intervista pilota è un passaggio fondamentale dello studio poiché è in grado di assicurare (in base alle difficoltà riscontare durante la stessa), per quanto possibile, che il questionario sia strutturato in maniera tale da ottenere le risposte necessarie ai fini conoscitivi. In definitiva, il questionario è articolato come riportato nel successivo Box 2.1.

L'analisi del contesto di riferimento risulta fondamentale nel processo di ideazione e sviluppo della cooperativa di comunità; in quanto, dalle testimonianze, dalle documentazioni e dal dibattito attuale si è appreso che le Cooperative di questo tipo si originano in zone contraddistinte da condizioni di vulnerabilità. La prima domanda del questionario si concentra sulle caratteristiche del comune, tentando di comprendere quali siano, per l'intervistato, le maggiori criticità riscontrate a livello sociale, economico e ambientale (con riferimento specifico alla agricoltura). Ci si aspetta una disamina delle problematiche emergenti relative agli avvenimenti che, più di altri, stanno caratterizzando negativamente le aree interne: isolamento, spopolamento, carenza di servizi essenziali e/o chiusura di attività commerciali (bar, alimentari, imprese ecc.), scarse opportunità di reddito per gli abitanti (disoccupazione di lunga durata e/o giovanile), marginalità sociale, presenza di aree dismesse e abbandonate, condizioni infrastrutturali e di dissesto idrogeologico in peggioramento (che impattano anche sui terreni agricoli), mancata valorizzazione del suolo, della viabilità agro-silvo-pastorale, caren-

za nei servizi di manutenzione (pulizia dei fondi, prevenzione boschiva, sistemazione idraulica ecc.).

Box 2.1. *Articolazione del questionario*

QUESTIONARIO PER LA RILEVAZIONE DELLE CRITICITÀ, DEI FABBISOGNI E DELLE MOTIVAZIONI DELLE COMUNITÀ LOCALI

1. Quali sono le criticità sociali, economiche e ambientali del territorio in cui vive?
2. Quali sono i fabbisogni comunitari?
3. Quali sono i fattori che impediscono di soddisfare i fabbisogni comunitari?
4. In che modo e con quali mezzi possono essere risolte le problematiche che impediscono di soddisfare i fabbisogni comunitari?
5. Vi sono soggetti (persone fisiche e giuridiche) capaci di generare consenso intorno a un progetto imprenditoriale?
6. Vi sono dei beni patrimoniali (immobili, superfici agrarie, boschi, ecc.) che potrebbero essere valorizzati e messi eventualmente a disposizione della comunità?
7. Vi sono dei servizi collettivi (piano neve, manutenzione del territorio, trasporto pubblico, mense scolastiche, ecc.) che potrebbero essere affidati alla cooperativa di comunità?

Nei comuni selezionati potrebbero essere presenti alcuni fenomeni descritti, per cui, appare chiaro, che per il buon funzionamento della società bisognerebbe assicurare ed erogare quei servizi che soddisfino i relativi bisogni e quegli interventi che rallentino e/o impediscano il reiterarsi di effetti negativi derivanti da queste stesse problematiche. La domanda inerente ai fabbisogni comunitari sorge spontanea, poiché le cooperative di comunità nascono per rispondere a un fabbisogno (inerente a una specifica zona) reale, percepito e condiviso da gran parte della collettività. Le domande successive, relative ai fattori che impediscono di soddisfare le esigenze riscontrate e quali possano essere i mezzi e modi di superamento degli stessi, sono necessarie per conoscere le motivazioni per cui i comuni finora non siano stati in grado di esperire le necessità comunitarie e per comprendere se questi ostacoli potrebbero essere di intralcio all'eventuale avvio di una cooperativa di comunità. Le risposte a tali quesiti forniscono informazioni utili anche per organizzare in maniera ottimale gli interventi e i servizi che la cooperativa di comunità dovrà erogare e per cogliere se vi siano delle risorse latenti o poco utilizzate su cui poter investire. La condizione imprescindibile per l'avvio di una cooperativa di comunità risulta però essere l'iniziativa collettiva, in quanto è una forma di innovazione sociale che si fonda sul capitale umano. La cooperativa di comunità si sta affermando sempre più come modello di cooperazione efficace poiché mette a sistema le attività di singoli cittadini, imprese, associazioni e istituzioni per il benessere e la sopravvivenza della comunità. La domanda inerente alla presenza di soggetti capaci di generare consenso intorno a un progetto imprenditoriale è stata formulata appositamente per valutare se, nei comuni oggetto di studio, ci sarebbero dei soggetti (in forma singola e associata) con la volontà di aggregarsi e collaborare per ricercare soluzioni ai fabbisogni comuni e produrre vantaggi a favore della comunità. Infine, gli ultimi due quesiti riguardano strettamente le opportunità e le ri-

sorse (latenti o sottoutilizzate) territoriali, sotto forma di beni patrimoniali e servizi collettivi, che potrebbero essere messi a disposizione e valorizzati da una futura cooperativa di comunità. Accanto all'iniziativa collettiva, nel processo generativo, l'impresa ha bisogno di un'infrastruttura sociale e relazionale a cui appoggiarsi per dare avvio alle proprie attività. Nella cooperativa di comunità i cittadini sono sia i produttori che i fruitori di beni e servizi e ogni Cooperativa si distingue per dimensioni, obiettivi, attività e servizi, poiché differenti sono le peculiarità e le condizioni della comunità, diversi i bisogni, le motivazioni e le modalità di risposta della collettività. Le proloco, le parrocchie, le associazioni, le imprese territoriali e in particolar modo la pubblica amministrazione incarnano quei soggetti di cui la Cooperativa necessita e con cui deve stringere relazioni soddisfacenti al fine di ottenere la gestione di determinati servizi all'interno del comune. La dimensione imprenditoriale rimane, comunque, la novità di questa concezione della cooperativa di comunità che deve essere intesa come una nuova forma di impresa che favorisce iniziative a scopo sociale in diversi ambiti (ambientale, turistico, agricolo ecc.). L'impresa si fonda su principi cooperativi per affrontare problemi sociali in un'ottica di presa in carico e valorizzazione dei beni comuni e di servizi collettivi, pubblici e di interesse generale, rispetto ai quali anche la pubblica amministrazione non è in grado di fornire risposte esaustive e soddisfacenti e con l'obiettivo principale di produrre beni e servizi che incidano sulla qualità della vita sociale ed economica, nonché sulla sopravvivenza stessa della comunità. Il questionario, così proposto, permette di far emergere quelle informazioni, relative ai territori selezionati, sostanziali per apprendere effettivamente se quelle zone siano contraddistinte da una serie di fattori che permetterebbero la nascita di una cooperativa di comunità. L'intervistatore ha cercato di far leva sui pareri e le opinioni di chi conosce appieno quei comuni, veicolando le loro conoscenze sugli elementi indispensabili per l'avvio di un progetto imprenditoriale. I testimoni privilegiati sono stati avvicinati in primis telematicamente, mediante l'invio di una lettera formale esplicativa dell'attività di ricerca che si stava conducendo e degli aspetti caratterizzanti le cooperative di comunità. La lettera presentava una duplice finalità: anticipava il tema dell'intervista ai futuri soggetti rispondenti in modo che non fossero del tutto sorpresi dalle domande che gli sarebbero state poste, e che riuscissero anche a essere più pronti e preparati nella risposta; valutava il grado di disponibilità dell'intervistato con l'inserzione dei contatti dell'intervistatore per un eventuale feedback (contattare l'intervistatore sarebbe stato sintomo di disponibilità e apertura nei confronti della specifica ricerca e delle cooperative di comunità). Il lavoro si pone l'obiettivo di condurre delle analisi sulla base delle interviste fatte ai sindaci e ad altri esperti dei comuni, che sono stati individuati nella fase di zonizzazione come vulnerabili. A ogni soggetto da intervistare è stato somministrato il questionario composto dalle sette domande precedentemente descritte con l'ausilio di un registratore, in modo tale che le risposte potessero essere riascoltate, elaborate e convertite in testo ai fini delle analisi. Le interviste sono state trascritte fedelmente,

senza modificare il modo di esporre dei soggetti coinvolti, senza correggere eventuali usi scorretti della lingua e/o refusi. Le trascrizioni dovevano essere accurate, a seguito della riproduzione di un pezzo o di tutta l'intervista bisognava riascoltare, inserire parti mancanti e correggere gli errori (sempre e solo quelli derivanti dalla trascrizione, e non quelli commessi dall'intervistato). L'importanza di effettuare una trascrizione accurata risiede nel poter applicare al meglio i metodi di analisi del testo che sono stati poi utilizzati: indici di leggibilità del testo, di seguito presentati a livello teorico; *Content* e *Sentiment Analysis* (oggetto del paragrafo seguente).

2.4. Gli indici di leggibilità

Gli indici di leggibilità sono delle formule matematiche che permettono di definire la difficoltà di lettura di un testo, in base a delle caratteristiche precise. Nella valutazione della leggibilità di un testo, però, non si tiene conto di una serie di fattori linguistici che possono ostacolarne o impedirne la comprensione; quest'ultima dipende dalla presenza o meno di parole difficili che possono rendere complesso, in termini di capacità d'intendere e giustificare il senso, il testo in questione. Per leggibilità, invece, si intende l'impianto linguistico di un testo che fa sì che lo stesso risulti più o meno chiaro e accessibile ai lettori, sulla base di un ampissimo numero di caratteristiche linguistiche in combinazione, a prescindere da quanto possa essere difficoltoso l'argomento trattato. Per creare una formula di un indice di leggibilità si può far riferimento ad alcuni parametri, quali fattori di leggibilità di un testo:

1. l'aspetto grafico: presenza di immagini, tabelle e disegni; disposizione del testo in capitoli, paragrafi e sotto-paragrafi e la titolazione di queste partizioni; l'utilizzo di caratteri speciali per segnalare definizioni e lessico;
2. la lunghezza delle frasi: tanto più una frase è lunga, e quindi ricca di subordinazioni, tanto meno sarà di facile e immediata lettura;
3. la lunghezza delle parole all'interno di ciascuna frase: quanto più una parola è lunga, tanto maggiore è il carico di informazioni che essa trasmette; la presenza di molte parole lunghe può rendere una frase troppo densa di significato e quindi di non facile lettura.

La leggibilità linguistica, pertanto, riguarda l'impiego della lingua considerando le sue diverse sfaccettature: scelta dei termini, della sintassi impiegata, articolazione dei contenuti, ecc. Sono state composte diverse formule per la leggibilità di un testo, ma quelle risultate migliori si basano su variabili linguistiche relativamente più semplici e di facile calcolo, come la lunghezza delle parole e la lunghezza delle frasi; variabili linguistiche che sono essenzialmente indipendenti dall'argomento e dal contenuto del testo. A tal proposito, la formula di leggibilità che ha ottenuto maggior successo e diffusione è nota come Formula di Flesch (prende il nome dall'inventore Rudolf Flesch), che considera per l'appunto solo due variabili linguistiche: lunghezza media delle parole espressa in sillabe per parola e la lunghezza media delle frasi espressa in

parole per frase. Secondo gli studi di Flesch un testo può essere definito difficile quando contiene molte subordinate (difficoltà sintattica) e parole astratte (difficoltà semantica): una parola lunga è usata generalmente meno di una breve, e una frase lunga, di solito, risulta più complessa – dal punto di vista sintattico – di una breve. La formula di Flesch, che deve la sua diffusione proprio alla semplicità, è nata per l'inglese ed è stata adattata alla lingua italiana da Roberto Vacca. L'indice di facilità di lettura di Flesch-Vacca si basa sulla seguente formula:

$$\textit{Facilità di lettura} = 206 - 0{,}65\,S - W.$$

in cui: la *S* rappresenta il numero di sillabe presenti in ogni 100 parole, la *W* è la media di parole per frase, 206 è la costante applicata per mantenere i valori tra 0 e 100 e 0,65 è la costante riferita alla lunghezza media delle parole italiane. I risultati della formula possono oscillare su una scala di valori compresi tra 0 e 100, dove lo 0 indica la leggibilità più bassa (testo di difficile lettura) e il 100 la leggibilità più alta (testo di facile lettura), con le classi di riferimento riportate in Tabella 2.2:

Tabella 2.2. *Classi di leggibilità di un testo (indice Flesch-Vacca)*

Valore	Difficoltà di lettura	Educazione scolastica
91-100	Molto semplice	Inferiore alla licenza elementare
81-90	Semplice	Licenza elementare
71-80	Abbastanza semplice	Inferiore alla licenza media
61-70	Normale	Licenza Media
51-60	Abbastanza difficile	Diploma di maturità
31-50	Difficile	Laurea breve
0-30	Molto difficile	Laurea e oltre

Nel 1982 il GULP (Gruppo universitario linguistico pedagogico, presso l'Istituto di Filosofia dell'Università degli studi di Roma «La Sapienza»), ha dato vita a una nuova formula partendo direttamente dalla lingua italiana, definendo l'indice di Gulpease. La formula Gulpease, a differenza di quella di Flesch-Vacca, si basa sul calcolo della lunghezza delle parole espresso in lettere, e non più in sillabe, semplificandone il calcolo automatico. L'indice di Gulpease è risultato essere il seguente:

$$\textit{Facilità di lettura} = 89 - LP/10 + FR \times 3$$

In cui: le costanti sono 89 e 10, le *LP* rappresentano le lettere incluse in 100 parole rapportate al totale delle parole del testo e le *FR* sono invece le frasi presenti in 100 parole, rapportare sempre sul totale delle parole. Per questo indice è stata prevista una scala d'interpretazione dei valori che, come per l'indice di Flesch, possono oscillare in una scala da 0 a 100. Il range però questa volta è ripartito in tre classi (Tabella 2.3):

Tabella 2.3. *Classi di leggibilità di un testo (indice Gulpease)*

Valore	Difficoltà di lettura	Educazione scolastica
> 80	Semplice per	Istruzione elementare
> 60	Semplice per	Istruzione media
> 40	Semplice per	Istruzione superiore

La scala mette in relazione i valori restituiti dalla formula con il grado di scolarizzazione del lettore: per esempio, un testo con indice Gulpease 60 è molto difficile per chi ha la licenza elementare, difficile per chi ha la licenza media, facile per chi ha un diploma superiore. Gli indici appena descritti saranno utilizzati, a seguito della trascrizione delle interviste, per prevedere se le informazioni trasmesse dai soggetti intervistati possono essere di facile comprensione, in termini di leggibilità.

Essendo gli intervistati soggetti definibili testimoni privilegiati, interpellati per il ruolo che ricoprono, ed essendo loro i conoscitori, più di altri, delle problematiche, ma anche delle risorse che contraddistinguono un determinato luogo, o meglio il territorio in cui vivono, risulta interessante comprendere se le loro risposte, quindi le informazioni che trasmettono (ognuno in maniera personale, secondo il proprio modo di esprimersi e di scegliere le parole) possano essere ritenute di un livello espressivo più o meno elevato e possano essere interpretate in maniera più o meno semplice. L'utilizzo degli indici appena presentati è sempre stato riservato alla comprensibilità di testi scritti; il loro impiego nell'ambito della valutazione del livello espressivo dei rispondenti a un'intervista non è certamente usuale e anzi, per quanto di nostra conoscenza, non risultano applicazioni di questo genere in letteratura.

2.5 I metodi di analisi testuale

L'indagine diretta attraverso cui è stato possibile rilevare le opinioni dei soggetti portatori di interesse, successivamente trascritte, ha consentito di mettere in atto una tipologia di analisi del tutto nuova per il settore di riferimento, che sta assumendo importanza crescente negli ultimi anni.

Si tratta della *Content Analysis*. Diverse sono le definizioni da poter attribuire a tale espressione. La più esaustiva è quella di Krippendorff (2013), che la definisce come «una tecnica di ricerca che consente di fare deduzioni replicabili da fonti testuali, sulla base del contesto a cui fanno riferimento» (Drisko, 2016). Nello specifico tale definizione fa riferimento al cosiddetto «contenuto manifesto» di un testo, ossia a ciò che è letteralmente presente in una comunicazione. I ricercatori fanno uso della suddetta tecnica di analisi per una molteplicità di scopi: identificare le attitudini degli individui o di gruppi di individui, conoscere i loro punti di vista, nonché le loro aspettative e interessi.

Si tratta quindi di un metodo di ricerca che rende possibile l'estrapolazione e l'analisi di informazioni utili contenute nei testi.

È possibile distinguere tre diversi approcci di Content Analysis (Drisko, 2016):

- *Basic Content Analysis:* tale espressione fa riferimento a un metodo di ricerca che conduce all'individuazione di aspetti oggettivi e quantitativi, espressi attraverso il contenuto manifesto di una comunicazione. Tale approccio utilizza quindi tecniche analitiche basate sulle frequenze di parole, attraverso le quali è possibile stabilire l'importanza di un determinato contenuto.
- *Interpretive Content Analysis:* Holsti (1969) descrive questo approccio come una procedura attraverso la quale è possibile fare inferenza, mediante l'oggettiva e sistematica individuazione di caratteristiche specifiche dei testi. La differenza rispetto al precedente approccio deriva dal fatto che quest'ultimo prende in considerazione sia il contenuto manifesto sia il contenuto latente. Per «contenuto latente» si intende il significato che non traspare in maniera immediata da una comunicazione, ossia quello implicito; pertanto è richiesta l'interpretazione del contenuto della comunicazione. È questo il caso di testi nei quali è possibile rinvenire artifici linguistici complessi, tra cui l'ironia, il sarcasmo, che richiedono anche una contestualizzazione del testo.
- *Qualitative Content Analysis:* si tratta di un approccio nato in Germania negli ultimi anni, così definito da Mayring: «approccio di analisi del testo, controllato, empirico, metodologico, che permette di analizzare il contenuto prescindendo dal contesto» (Mayring, 2000). Si tratta quindi di un approccio che favorisce l'analisi di testi di vario tipo, mediante l'analisi sia del contenuto manifesto, sia di idee centrali che rappresentano il contenuto primario di un testo. Infatti, se inizialmente la Content Analysis nasceva soltanto come uno strumento per analizzare testi scritti; attualmente lo sviluppo di nuovi mezzi di comunicazione ha reso possibile la sua applicazione a fonti diverse. Per i ricercatori con la parola «testo» si fa riferimento a un'ampia gamma di mezzi di comunicazione, quali registrazioni audio, video, immagini ecc. È infatti possibile trasformare tali fonti in testo, ricorrendo alla trascrizione, operazione che comporta però la perdita di alcune informazioni legate alla forma originale del messaggio, come il tono/ritmo di voce nel caso delle interviste. Spesso i data sets oggetto di questa tipologia di analisi sono rappresentati da interviste. È proprio questo l'approccio adottato nel presente lavoro.

La Content Analysis presenta tuttavia vantaggi e svantaggi. Le problematiche più comuni sono legate alle difficoltà che si riscontrano dall'estrapolazione delle informazioni da contesti tra loro eterogenei. D'altro canto, il principale beneficio della tecnica è proprio dato dalla possibilità di poter eseguire la stessa su dati non strutturati. Al fine di condurre la Content Analysis è stato necessario fare ricorso alle tecniche di data mining[3] e di text analysis.

[3] Con l'espressione «data mining» si indica «il processo di ottenimento di conoscenze utili da insiemi di dati di grandi dimensioni, mediante l'impiego, in maniera automatica o semiautomatica, di tecniche informatiche e statistiche» (Zani e Cerioli, 2007).

L'espressione *Text Mining* o *Text analysis* racchiude in sé una pluralità di significati, che hanno come denominatore comune l'utilizzo di testo come input dal quale estrapolare informazioni che rendono possibile analisi di vario tipo. La text analysis presenta diversi campi di applicazione. In campo medico, per esempio, lo scienziato Swanson ha dimostrato come l'utilizzo di informazioni contenute nei testi di letteratura, possono rivelarsi utili per formulare ipotesi circa le cause che determinano l'insorgere di malattie rare (Swanson, 1986); in campo economico, la text analysis consente di analizzare i livelli di customer satisfaction e la customer retention; la text analysis in campo sociale può essere uno strumento per fare previsioni su determinati accadimenti futuri.

Nel lavoro eseguito l'attenzione si è focalizzata sullo studio e sull'individuazione, mediante questa metodologia, delle criticità territoriali, delle motivazioni alla base della costituzione delle cooperative di comunità e della definizione del loro ruolo a supporto dei fabbisogni comunitari.

L'idea alla base della text analysis è quella di trasformare il testo in un formato strutturato, costituito da dati espressi sotto forma di frequenze su cui applicare le tradizionali regole di data mining. Sono numerosi i metodi che negli ultimi anni sono stati utilizzati per il raggiungimento di questo scopo e con la crescente importanza assunta dalla text analysis nella ricerca sulla comunicazione, molti ricercatori fanno affidamento sull'uso di software avanzati che rendono possibile tale analisi. Il software adoperato per le elaborazioni è R. L'esecuzione della text analysis è avvenuta secondo le regole standard, ossia si è svolta seguendo una serie di step di seguito elencati, che vanno dalla preparazione dei dati fino alla loro analisi (Welbers, 2017).

1. La fase di preparazione dati si sviluppa a sua volta nelle seguenti fasi: importazione del testo, preprocessing e creazione della document term-matrix (dtm).
   - L'importazione del testo ha permesso, appunto, di importare il testo delle registrazioni a disposizione nel software utilizzato per l'esecuzione della text analysis.
   - La fase di preprocessing è quella che è risultata più articolata, in quanto racchiude in sé una serie di operazioni che consentono di pulire il testo importato da elementi non significativi ai fini dell'analisi. Attraverso la tokenizzazione è stato possibile suddividere il testo in «tokens», ossia in parole che costituiscono l'elemento chiave per l'estrapolazione della componente semantica. Si tratta di uno step che non risulta di facile realizzazione, soprattutto nei casi in cui le parole del testo non sono separate da spazi bianchi. Attraverso la normalizzazione è stata attuata una trasformazione delle parole in una forma più uniforme. Nello specifico i vantaggi principali che la suddetta operazione apporta a una analisi del testo sono: da un lato la possibilità di individuare parole che presentano lo stesso significato e dall'altro la possibilità di andare a ridurre le dimensioni del vocabolario. Una tecnica di normalizzazione importante che consente al software di riconoscere se due o più parole sono tra loro identiche, consiste nel trasformare tutto il testo oggetto di analisi in lettere minuscole. Si parla

in tal caso di lowercasing. Nell'esecuzione di un'analisi del testo bisogna prendere in considerazione anche il fatto che una stessa parola può presentare diverse variazioni morfologiche; questo avviene sia per le coniugazioni verbali (ad esempio "mangiare" e "mangio"), sia nel caso in cui la parola viene espressa al plurale (per esempio "aiuto" e "aiuti"). Quindi vi sono termini che si caratterizzano per avere una stretta relazione semantica, in quanto presentano una forma base standard o stessa radice e dei suffissi che possono variare. Per far fronte a queste situazioni è stata messa in pratica un'altra tecnica di normalizzazione, detta stemming. Quest'ultima consente di individuare le parole che presentano la stessa radice e di riportarle alla loro forma base depurate dai suffissi che possono variare, così da poter avere una riduzione dei termini presenti nel testo.
Infine l'ultima operazione facente parte del preprocessing è la seguente: rimozione di stop words. In un testo vi sono parole ricche di significato il cui contributo può essere determinante per lo svolgimento di una text analysis, ma vi sono anche parole che non forniscono informazioni di rilievo sul contenuto di un testo. È questo il caso, ad esempio, degli articoli. Andare a porre un filtro per tali parole, quindi andare a eliminarle durante il processo di analisi, è stato necessario, non solo al fine di ridurre la dimensione del testo, ma anche al fine di rendere l'analisi più accurata ed efficace e di ridurre il carico di calcolo. Per rimuovere tali parole, è stato necessario collegarle a liste predefinite di stop words. Oltre alla rimozione delle stop words sono stati eliminati dal testo, attraverso opportuni comandi, anche i caratteri numerici e i segni di punteggiatura, poiché irrilevanti (Welbers, 2017).

– Sulla base dei risultati ottenuti dalle operazioni precedenti, attuate per ciascuna risposta data dai soggetti intervistati, è stato possibile realizzare le matrici «DTM». La Document-term matrix costituisce uno dei formati più comuni per la rappresentazione di un corpo di testo o *corpus*, dove tale espressione sta a indicare un insieme strutturato di testi in un formato del tipo «*bag-of-words*». Si tratta quindi di un modo attraverso il quale è possibile inserire il testo in una matrice, le cui righe rappresentano i documenti, le colonne i termini e ciascuna cella indica la frequenza con cui ogni termine si presenta in ciascun documento (Munzert, 2015). Pertanto la DTM appare come una rappresentazione che offre il vantaggio di poter lavorare con matrici e vettori, quindi sposta l'attenzione dal testo ai numeri, che risultano più semplici da analizzare.

2. Conclusa la fase di preparazione dei dati è stata effettuata la loro analisi. I celebri insegnanti propongono diversi approcci di analisi (Boumans e Trilling, 2016):
   – counting and dictionary;
   – supervised machine learning;
   – unsupervised machine learning;
   – statistical.

L'approccio counting and dictionary si caratterizza per l'utilizzo di modelli diversi, come ad esempio parole, *query Booleane* e *regular expressions*[4], che consentono di contare la frequenza con la quale alcuni concetti si presentano nel testo. Tale metodo si caratterizza per l'utilizzo di particolari «dizionari», uno strumento semplice da utilizzare, che riesce ad apportare diversi vantaggi nella conduzione di analisi del testo. Tale approccio è di tipo deduttivo, cioè è basato su precodifica. Il dizionario consente quindi di definire a priori i codici che sono oggetto di misurazione e in che maniera questa viene svolta.

Con l'espressione *Supervised machine learning* si fa riferimento a un metodo di apprendimento automatico, basato su un insieme di tecniche che favoriscono la costruzione di classificatori di testo mediante particolari algoritmi che codificano il testo oggetto di analisi sulla base di esempi di codifica, i cosiddetti «dati di training o di addestramento», a esso forniti. L'esempio che meglio esemplifica tale approccio è dato dalla Sentiment Analysis. Si tratta di un approccio che può essere considerato sia deduttivo sia induttivo. Deduttivo perché gli algoritmi lavorano prendendo a riferimento degli esempi precostituiti creati da ricercatori, induttivo perché i ricercatori non forniscono le regole per la ricerca dei codici.

È opportuno precisare che i dati precodificati presi come riferimento per l'esecuzione dell'indagine possono essere anche soggetti a errori e questo può rivelarsi un punto di debolezza per l'analisi, i cui risultati potrebbero essere non perfettamente attendibili.

Esistono tre modelli di apprendimento supervisionato:

– *Support Vector Machines*: tale modello impropriamente detto «vettore di supporto» è uno dei più conosciuti. Si caratterizza per l'impiego di una rappresentazione spaziale dei dati. Nello specifico può essere pensato come una superficie che rappresenta il confine tra diversi punti di dati, che costituiscono esempi tracciati nello spazio multidimensionale sulla base dei loro valori di funzionalità. Lo scopo che si vuole raggiungere mediante l'utilizzo di tale modello è quello di creare un iperpiano, tecnicamente definito «*hyperplane*»[5], che consenta la partizione di dati che siano omogenei su entrambi i lati, creando così gruppi di dati formati da elementi simili tra loro (Lantz, 2015).

[4] Le *regular expressions* o espressioni regolari sono dei modelli di testo generalizzabili che vengono utilizzate per la ricerca e per la manipolazione di dati all'interno di un corpo di testo. Vengono adoperate nell'ambito dell'approccio counting and dictionary in quanto sono convenzionali. La potenzialità nell'utilizzo di tali strumenti sta nella possibilità di rendere le query di ricerca più flessibili e generalizzate (Munzert, 2015).

[5] Un iperpiano può essere definito come una superficie piatta in uno spazio di dimensioni elevate. Tradizionalmente si tende a rappresentarlo per semplicità come una linea nello spazio bidimensionale, a causa della complessità di rappresentazione a cui si dovrebbe far fronte per operare in uno spazio che presenta dimensione superiore a due (Lantz, 2015).

– *Random Forest*: è questo un modello che consente la creazione di più alberi decisionali. Un albero decisionale si compone di diversi livelli che richiedono consecutivamente se una determinata caratteristica è presente o meno in un documento. A seconda della presenza o assenza della stessa viene presa una decisione. Nello specifico le decisioni vengono prese prendendo in considerazione le frequenze osservate di presenza o assenza di funzioni nel set di dati di addestramento. Nel modello Random Forest poiché gli alberi decisionali sono molteplici, le previsioni vengono effettuate sulla base della frequenza osservata più alta nei diversi alberi (Munzert, 2015).
– *Maximum Entropy*: è un modello che rispecchia il modello logit multinomiale. Si cerca attraverso quest'ultimo di stimare l'appartenenza in sei diverse categorie d'attualità (Munzert, 2015).

L'*Unsupervised machine learning* rappresenta un'alternativa alle tecniche di apprendimento supervisionato per la classificazione del testo. Tale approccio non richiede l'utilizzo di dati di addestramento per effettuare la categorizzazione del testo e non vengono specificate regole di codifica. L'unica influenza che il ricercatore può esercitare è quella di specificare alcuni parametri, come il numero di categorie in cui i documenti sono classificati, di conseguenza diversi sono gli svantaggi a esso riconducibili. Innanzitutto i ricercatori difficilmente riescono a determinare e specificare i limiti di uno schema di categorizzazione; in secondo luogo la difficoltà emerge nel momento in cui bisogna interpretare i risultati, dato che l'analisi non presenta un contesto di riferimento (Welbers, 2017). Per sopperire a tali svantaggi, si può comunque ricorrere all'utilizzo nello stesso caso sia del metodo di apprendimento supervisionato sia del metodo di apprendimento non supervisionato, cosicché possano completarsi a vicenda. Non sono infatti metodi tra loro concorrenti, come è stato sostenuto da Grimmer e Stewart (2013), in quanto consentono di assolvere a scopi diversi. Nel caso in cui i documenti devono essere inseriti in categorie predeterminate, è l'approccio di apprendimento supervisionato il più adatto, per il fatto che un approccio non supervisionato non riuscirebbe né a determinare categorie appropriate al caso né a interpretarle in maniera adeguata. L'approccio non supervisionato non presenta però solo limiti; esso infatti può determinare l'importante vantaggio di fornire categorie che i ricercatori possono non aver considerato.

Le principali tipologie di metodi di apprendimento non supervisionato sono le seguenti:

– *Topic Model*: è questo un modello basato su una tecnica detta *«Latent Dirichlet Allocation (LDA)»*. Il modello assume che il corpo di testo di ciascun documento è costituito da un insieme di argomenti, detti appunto *topics* e che a ogni termine del documento possa essere assegnata una certa probabilità con riferimento all'appartenenza del termine all'argomento. Il numero di topics in cui il testo deve essere suddiviso può essere stabilito in maniera arbitraria. Uno dei punti di debolezza di tale modello deriva

dal fatto che non permette di prendere in considerazione le relazioni che possono esserci tra i diversi topics (Munzert, 2015).

– *Cluster Analysis* di cui si è ampiamente parlato precedentemente al Par. 2.2.

L'ultimo approccio di analisi si incentra su tecniche statistiche e assume particolare rilievo in una text analysis. Un corpo di testo può essere infatti descritto, esplorato e analizzato mediante numerose tecniche statistiche. Una tra queste che è divenuta molto popolare, consiste nell'estrapolare il valore informativo che ciascun termine presenta all'interno del testo e nel classificarlo, al fine di visualizzare le parole che racchiudono maggiori informazioni come una nuvola di parola, la cosiddetta *word cloud* (Welbers, 2017). Si ottiene quindi una rappresentazione visiva, nella quale le parole che presentano un font di dimensione più grande sono quelle che si considerano più importanti, quelle che appaiono più piccole risultano invece meno importanti e informative all'interno del testo. Altre tecniche statistiche molto utilizzate sono quelle che consentono di operare con concetti quali la leggibilità o la diversità lessicale e prendono in considerazione sia la lunghezza della frase sia il numero di parole e di sillabe presenti in un testo.

Le tecniche di preparazione dei dati e di analisi delle parole costituenti un corpo di testo appena esposte, costituiscono soltanto il punto di partenza per l'esecuzione di una text analysis. Talvolta per condurre particolari tipi di analisi è necessario fare ricorso a tecniche avanzate di analisi, che prevedono l'utilizzo di software esterni. Per questo risultano più difficili da mettere in atto e richiedono una maggiore attenzione. Si parla in tal caso di *advanced natural language processing*[6] (Welbers, 2017).

Nel presente lavoro è stato seguito sia l'approccio di analisi di tipo supervisionato, sia l'approccio di tipo statistico. È stata dapprima eseguita una Sentiment Analysis, ossia una metodologia di analisi che favorisce l'estrazione e l'analisi del sentimento, delle emozioni, delle opinioni che sono espresse nei testi e che possono riguardare, ad esempio, un particolare brand, un prodotto, un servizio, un evento, un argomento ecc. (Liu, 2010). È quindi un concetto che fa riferi-

[6] Le principali tecniche di analisi avanzata sono:
Lemmatization: tecnica molto simile a quella di stemming. Consente di ricondurre le parole alla propria forma base, non andando a tagliare la parte finale della parola, bensì mediante l'utilizzo di appositi dizionari che consentono di sostituire la parola con il proprio lemma, ottenendo risultati più precisi.
Named Entity Recognition: tecnica che consente di verificare se una certa parola o anche una sequenza di parole, possano identificare un'entità e consente anche di determinarne il tipo; ad esempio si può verificare se una data entità identifica una persona, un'organizzazione o un'entità di altro tipo.
Part-of-Speech Tagging: tecnica molto utilizzata nei casi in cui si manifesta la necessità di filtrare determinate parole, per focalizzare l'attenzione e l'analisi su determinate categorie grammaticali, ad esempio articoli, pronomi, al fine di studiare eventi simili e meglio comprendere il linguaggio soggettivo. Per POS tags si intendono infatti categorie morfo-sintattiche per le parole, ad esempio nomi, verbi, aggettivi.

mento all'analisi del testo come strumento attraverso il quale è possibile identificare informazioni soggettive dalle fonti a disposizione. Nel momento in cui ci si trova a prendere decisioni, di qualunque tipo esse siano, che riguardino il singolo individuo o più in generale le organizzazioni, può risultare importante conoscere le opinioni altrui, quello che gli altri pensano, ed è proprio in queste situazioni che entra in gioco la Sentiment Analysis, un valido strumento di aiuto. Spesso si utilizza anche l'espressione *opinion mining* per fare riferimento a tale tipologia di analisi. È un campo di ricerca la cui importanza sta crescendo soprattutto nell'ambito del text mining e della Content Analysis[7]. I suoi campi di applicazione sono molteplici: politico, sociale, medico, marketing. Nello specifico il campo nel quale la Sentiment Analysis viene maggiormente utilizzata, è quello aziendale, perché consente di individuare i punti di forza e di debolezza di un'azienda, ma anche dei prodotti e servizi che questa offre e più in generale di un particolare brand. Si tratta di aspetti che non vanno sottovalutati e che aiutano a comprendere sia se un'azienda sta mantenendo o meno una posizione di vantaggio competitivo nel suo mercato di riferimento, sia se possono essere evitate crisi di corporate reputation[8], sia se è necessario migliorare le strategie e la pianificazione di marketing al fine di soddisfare le esigenze degli stakeholder. Nel presente studio si fa appunto riferimento alle cooperative di comunità e ai servizi che le stesse possono offrire nelle aree sottoposte a indagine diretta.

Nella conduzione di una Sentiment Analysis è possibile seguire tre tipologie di approcci: rilevamento di *keywords*, metodo *lexicon based* o delle affinità lessicali e metodi statistici.

Il metodo adottato nella sua esecuzione in questa ricerca, è definito *"lexicon based"*, detto anche di classificazione. Si tratta di un metodo che non solo consente di individuare nel testo delle keywords significative, dotate di intensità semantica, ma consente anche di «assegnare a parole arbitrarie un'affinità probabile a emozioni particolari». Ciò sta a significare che a ciascuna parola si assegna una polarità, un *"orientamento semantico"*, che non è altro che una misura della forza della parola all'interno del testo (Taboada, 2011). I concetti di Sentiment Analysis e orientamento semantico vengono spesso confusi. Si tratta

[7] Di recente la Sentiment Analysis sta assumendo rilievo crescente grazie all'evoluzione che si è registrata del World Wide Web. Il Web ha infatti determinato un cambiamento radicale del modo in cui le persone esprimono le proprie opinioni, principalmente mediante recensioni online che si rivelano utili non solo per il singolo individuo, ma anche per le organizzazioni e le aziende che in passato dovevano ricorrere a questionari, sondaggi e consulenze per conoscere le opinioni dei consumatori sui prodotti dell'azienda stessa, ma anche sui prodotti offerti dalla concorrenza. Al giorno d'oggi l'importanza che questa tipologia di analisi sta assumendo, è rafforzata dall'enorme crescita che sta interessando i social media, i blog, i forum, i social network, all'interno dei quali è possibile rinvenire una grossa mole di dati presenti in formato digitale dai quali estrapolare informazioni di vario genere (Liu, 2010).

[8] La *corporate reputation* è la considerazione di cui gode un'organizzazione in virtù della sua capacità di soddisfare le aspettative degli stakeholder nel tempo. Esprime il giudizio dei vari pubblici sull'azienda, confermato dalle esperienze dirette degli stakeholder e dalle azioni e dai risultati passati dell'organizzazione (<www.glossariomarketing.it>).

di espressioni strettamente connesse, tra loro interdipendenti. Qual è la differenza? Come già affermato, la Sentiment Analysis è il metodo che consente di estrapolare e analizzare un sentimento/opinione racchiuso in un testo, l'orientamento semantico è una misura della soggettività di un testo, si riferisce quindi alla forza delle parole, alla loro polarità. Si può quindi ribadire che la Sentiment Analysis rappresenta il metodo che consente di estrapolare e analizzare l'orientamento semantico (Taboada, 2011). Il metodo lexicon based prevede l'utilizzo dei cosiddetti «dizionari», ossia particolari contenitori al cui interno sono racchiusi indicatori dell'orientamento semantico del testo, solitamente rappresentati da aggettivi, sostantivi, avverbi e verbi, la cui costruzione può avvenire in maniera manuale, semi-automatica o in maniera del tutto automatica[9]. Il modo che si sceglie di seguire per la creazione del dizionario è importante, in quanto influisce sull'accuratezza dei risultati che si ottengono dall'analisi. I dizionari generati automaticamente risultano meno precisi e meno stabili rispetto alle altre due tipologie. Infatti in quest'ultimo caso le piattaforme utilizzate per l'associazione della polarità ai commenti non riescono a prendere in considerazione concetti emotivi complessi quali l'ironia. Questo significa che se attraverso un commento ironico si vuol esprimere un concetto negativo, al contrario gli verrà attribuita una polarità positiva. Si deve quindi mettere in evidenza il fatto che non sempre le opinioni vengono espresse attraverso l'utilizzo di *opinion words*; spesso si fa ricorso ad artifici linguistici quali appunto le figure retoriche, nonché all'utilizzo di espressioni non formali, *slang* e *emoticons* che consentono di dare maggiore enfasi, espressività all'opinione. Altra problematica che l'utilizzo dei dizionari creati in maniera automatica presenta, è legata al fatto che spesso le parole utilizzate possono assumere un significato differente a seconda del contesto al quale si riferiscono, per cui può rivelarsi sbagliato prendere in considerazione il significato semantico dei singoli tokens. La maggior parte della ricerca statistica sulla classificazione del testo crea classificatori di testo del tipo Support Vector Machine (SVM), costruiti sulla base di specifici set di dati costituiti principalmente da unigrams o bigrams[10] (Taboada, 2011). Come già spiegato precedentemente, tali classificatori, poiché rientrano nella tipologia di apprendimento supervisionato, risultano più adatti nel rilevamento e nell'estrazione della polarità dei testi e conducono a risultati più accurati e precisi. Bisogna però precisare che tali performance risultano elevate nel caso

[9] Nei primi tempi in cui la Sentiment Analysis ha iniziato ad assumere importanza, le ricerche erano incentrate sulla ricerca del sentiment dagli aggettivi, in quanto venivano considerati come parole che in misura maggiore rispetto alle altre racchiudevano contenuto soggettivo all'interno di un testo. Soltanto recentemente l'attenzione si è estesa anche all'analisi di verbi, nomi e avverbi, in quanto si è fatta strada la concezione secondo cui l'orientamento semantico di un intero documento è l'effetto della combinazione delle parole che formano un testo, in quanto ciascuna può essere espressione di soggettività (Taboada, 2011).

[10] Con il termine unigrams si fa riferimento ai tokens presi singolarmente, mentre con il termine bigrams ci si riferisce ai tokens presi a coppie. Generalizzando si parla di *n*-grams per far riferimento a gruppi costituiti da *n* parole (Welbers, 2017).

in cui il classificatore viene adoperato nel suo dominio specifico, ossia quello rispondente alla tipologia di analisi per il quale è stato creato, quindi richiede che vi sia coerenza tra l'analisi che deve essere condotta e la tipologia di dati di addestramento utilizzati.

Nel presente lavoro i dizionari sono stati costruiti manualmente. Sono stati seguiti i seguenti passaggi: per ciascuna zona oggetto di analisi, Cratere, Trigno e Fortore, e conseguentemente per ciascuna risposta alle domande del questionario, sono state prese in considerazione tutte le parole risultanti dalle matrici DTM e sono state riportate in un file excel (per risposta, questo significa risposta 1, risposta 2,…, risposta 7). A ciascuna di esse è stato attribuito un orientamento semantico, sulla base di una scala che va da (polarità fortemente negativa) a (polarità fortemente positiva). Alle parole neutre è stato attribuito un orientamento semantico pari a 0. La positività o la negatività sono state stabilite sulla base del significato che la parola assume nella maggior parte dei contesti. Quindi per ciascuna area sono stati in questa maniera costruiti sette dizionari importanti in R, uno per domanda, utilizzati nella determinazione del sentiment complessivo di ciascuna risposta data dagli intervistati. I dizionari sono il risultato del lavoro di più ricercatori, in quanto è necessario garantire la veridicità dei risultati legati all'analisi in oggetto.

Quando si conduce una Sentiment Analysis bisogna prestare attenzione anche ai cosiddetti intensificatori, ossia parole che non possono essere prese in considerazione separatamente rispetto a quelle adiacenti in quanto ne rafforzano o meno l'intensità semantica. Gli intensificatori a seconda della loro polarità, possono essere classificati in due categorie (Taboada, 2011):

- amplificatori: svolgono la funzione di andare ad accrescere l'intensità semantica della parola alla quale si riferiscono, quindi sono positivi. L'amplificatore più utilizzato è rappresentato dalle parole "molto", "più", ecc.
- downtoners: svolgono la funzione di andare a ridurre l'intensità semantica della parola adiacente. Sono quindi negativi. Un esempio di downtoner è rappresentato dalle parole "meno", "leggermente", "poco", ecc.

Gli intensificatori non sono gli unici elementi in grado di modificare la polarità della parola alla quale si riferiscono. Nello svolgimento di una Sentiment Analysis bisogna prendere in esame anche la presenza di negazioni. La funzione svolta dalla negazione è semplicemente quella di invertire la polarità della parola a cui si riferiscono, da positiva a negativa e viceversa. L'esempio emblematico di negazione è dato dall'avverbio "non".

Pertanto, al fine di tener conto degli aspetti appena evidenziati e della contestualizzazione, è stato altresì costruito un ulteriore dizionario, nel quale sono stati inseriti sia gli intensificatori sia le negazioni. Per determinare l'orientamento semantico complessivo di ciascuna risposta è stato innanzitutto necessario utilizzare dei «modificatori», che associano delle probabilità a ogni parola intensificante, modellando appunto l'intensificazione. Proprio per il fatto che gli intensificatori vengono implementati attraverso dei modificatori, si può affer-

mare che questi riescono a catturare in maniera accurata la varietà di parole che vanno a intensificare e a esprimerne l'orientamento semantico.

Successivamente dopo aver effettuato in R, confronti tra le parole presenti all'interno di ciascun dizionario e le parole racchiuse all'interno di ciascuna matrice DTM ottenuta per risposta, è stato determinato il sentiment complessivo di queste ultime nel seguente modo:

$$Sent.parziale = n(intens;neg) \times S.O.(intens;neg) \times S.O.(word)$$

dove $n(intens;neg)$ è il numero di volte che l'intensificatore o la negazione è legata alla parola presa in esame; $S.O.(intens;neg)$ è l'orientamento semantico associato alle parole intensificanti o ai negatori e $S.O.(word)$ è l'orientamento semantico associato alla parola esaminata.

$$Sent.tot. = Sent.parz. + [n_{tot} - n(intens;neg)] * S.O.(word)$$

dove n_{tot} rappresenta la frequenza con cui la parola in oggetto appare nel testo.

Infine è stato messo in atto l'approccio di analisi di tipo statistico, attraverso il quale è stato possibile rappresentare delle word cloud per ciascuna risposta, che mettono in evidenza le parole che presentano una forza maggiore all'interno del testo, forza determinata non solo dalla frequenza con cui esse vengono rilevate, ma anche dall'orientamento semantico calcolato per ciascuna di esse.

2.6 L'analisi economico-aziendale

L'analisi economico-aziendale ha lo scopo di valutare la fattibilità degli elementi di definizione del modello di cooperativa di comunità. Essa viene condotta su un'area bersaglio individuata nel corso della ricerca, in considerazione delle condizioni di vulnerabilità, delle motivazioni, dei soggetti capaci di generare consenso intorno a un progetto imprenditoriale, delle risorse dormienti e dei servizi pubblici che possono essere affidati alle cooperative di comunità.

Lo studio di scenario viene eseguito attraverso un approccio economico-aziendale che mette in luce le caratteristiche patrimoniali, economiche e di redditività di quattro cooperative di comunità, selezionate nella banca dati AIDA[11], che operano negli stessi ambiti di attività economica dell'ipotetica costituenda cooperativa di comunità. Le cooperative verranno indicate con le lettere A, B, C, D e sono state costituite rispettivamente nel 1991, 2003, 1984, 1952.

Gli ambiti di attività sono i seguenti: agricoltura, selvicoltura, agroalimentare, servizi ambientali e turismo.

I dati presi in considerazione attengono all'assetto patrimoniale, a quello economico e alla redditività.

[11] La banca dati AIDA contiene informazioni sulle società operanti in Italia. I dati economici e patrimoniali che fornisce su ogni singola azienda sono: il bilancio, il settore di attività economica, le informazioni anagrafiche, il numero dei dipendenti, le unità locali, l'azionariato e le partecipazioni. La banca dati contiene anche informazioni descrittive come l'anno di costituzione e la forma giuridica.

Per quanto riguarda l'assetto patrimoniale sono utilizzate le seguenti informazioni:
1. Attivo;
2. Patrimonio netto (PN);
3. Capitale sociale (CS);
4. Debiti totali;
5. Debiti su fatturato.

La voce relativa all'Attivo rappresenta l'ammontare complessivo degli investimenti realizzati dalle cooperative. Il patrimonio netto o capitale netto rappresenta la fonte di finanziamento interna dell'azienda. Esso rappresenta la differenza tra attività e passività e l'insieme delle risorse di cui l'azienda dispone al suo interno. È quindi l'insieme delle disponibilità finanziarie introdotte, in modo diretto o indiretto, dalla proprietà per lo svolgimento delle attività dell'impresa. Esso è rappresentato da Capitale sociale, Riserve, Utili da destinare o Perdite in sospeso. Il Capitale sociale, chiamato anche capitale di rischio, è composto dalle risorse immesse dai soci al momento della costituzione della società. Svolge il ruolo di protezione della proprietà dal rischio di fallimento e ha la funzione di garanzia per il rimborso dei creditori. Costituisce quindi una sorta di «somma fissa» per il soddisfacimento dei debiti contratti dall'impresa. L'ammontare del capitale sociale può variare in aumento o in diminuzione. L'aumento del capitale sociale può essere deciso dai soci che per bilanciare tale aumento dovranno emettere nuove quote. Diversamente la diminuzione di capitale sociale può essere eseguita in particolari situazioni economiche e patrimoniali.

I debiti totali rappresentano il complesso delle passività contratte dalle imprese per coprire il proprio fabbisogno finanziario. Analizzare la consistenza e la tipologia dei debiti detenuti è un elemento fondamentale per progettare strategie di crescita e di sviluppo mirate.

Il peso dei debiti sul fatturato è un valore percentuale che viene calcolato annualmente come il rapporto tra i debiti totali sui ricavi totali di ogni singola impresa.

Gli aspetti patrimoniali analizzati per lo studio di scenario sono costituiti dagli elementi presi dai bilanci d'esercizio nelle sezioni dell'attivo (dove sono indicate le risorse disponibili e il loro impiego) e nelle sezioni relative alle passività (dove vengono ascritte le fonti di reddito che includono anche il capitale dei soci). Inoltre per un quadro completo relativo agli impieghi viene calcolato, in termini percentuali, anche il peso dei debiti rispetto al fatturato e il peso dei debiti contratti con le banche, anch'essi rispetto al fatturato.

Relativamente all'aspetto economico le informazioni utilizzate sono le seguenti:
1. Costi totali;
2. Costo del lavoro;
3. Costo del lavoro pro capite;
4. Ricavi totali;
5. Valore aggiunto totale;
6. Valore aggiunto su attivo.

I costi totali della produzione rappresentano l'ammontare totale degli oneri legati dell'attività produttiva. Si ottengono sommando singolarmente tutti i costi di produzione. Il costo totale può essere inteso anche come l'esborso economico realizzato per l'acquisto dei fattori produttivi e dei beni funzionali all'attività d'impresa come: macchinari, impianti, fabbricati e di tutti gli strumenti necessari a garantire il corretto funzionamento del processo produttivo.

La voce del costo del lavoro rappresenta l'ammontare totale delle spese realizzate da un'impresa per remunerare la propria forza lavoro. Normalmente rappresenta la parte più consistente dei costi di produzione delle imprese. Il costo del lavoro si compone della somma di diverse voci che vanno a determinare il salario netto che viene erogato ai lavoratori. Esso è composto dal salario-stipendio lordo del lavoratore che viene scorporato in due tipologie di quote. La prima è la quota a carico del dipendente (comprendente imposte, contributi sociali e contributi per assicurazioni obbligatorie) e la seconda è la quota a carico dell'imprenditore (rappresentata dall'insieme di contributi sociali, assicurazioni obbligatorie comprendenti ratei di tredicesima mensilità aggiunte delle altre mensilità, ratei del TFR, ferie e permessi maturati e ogni altro importo attinente alla prestazione lavorativa da conteggiare a consuntivo vista la natura non prevedibile).

È utile evidenziare, in ogni caso, che le cooperative di comunità potrebbero ricorrere a forme di lavoro volontario e/o mutualistico al fine di contenere il costo del lavoro.

Il costo del lavoro pro-capite è il rapporto tra costo del lavoro e numero medio di dipendenti assunti. È un indice che permette di quantificare la spesa che le aziende sostengono annualmente per ogni lavoratore.

I ricavi totali esprimono il valore monetario totale delle entrate.

Il valore aggiunto è l'incremento di valore che l'azienda riesce a ottenere nella produzione e distribuzione di beni e di servizi finali partendo dalle risorse iniziali immesse nel ciclo produttivo. Si ottiene dalla differenza fra il valore totale della produzione (riferito ai beni e ai servizi prodotti) e i costi realizzati per produrre ogni singola unità produttiva. Questa voce indica quanto peso possiedono i fattori produttivi interni delle aziende (come capitale, lavoro e tecnologia) rispetto ai fattori produttivi acquistati esternamente, per ottenere un dato livello di produzione. La ricchezza prodotta dalle cooperative viene ripartita tra tutti i soggetti che hanno partecipato al processo produttivo.

Il valore aggiunto su attivo è un indice percentuale che si ottiene dal rapporto tra il valore aggiunto riferito a ogni singolo anno e il rispettivo valore degli investimenti. La performance del rapporto dipende dall'ammontare complessivo degli investimenti realizzati annualmente dalle imprese. Maggiore è la quota di investimenti strutturali (dipendenti, beni mobili e beni immobili) che realizzano, più grande è la capacità di utilizzare i fattori produttivi interni dell'azienda.

La sezione relativa agli aspetti economici delle cooperative è costituita dall'insieme di:

- elementi di natura contabile selezionati all'interno dei Conti Economici (nelle sezioni relative a «Valore della produzione» e «Costi della produzione»;

- indici economici volumetrici calcolati in termini e valori assoluti o in tassi percentuali.

La redditività è stata calcolata attraverso i seguenti indicatori:

1. Return on investment (ROI);
2. Return on equity (ROE);
3. Return on sales (ROS).

Gli indici di bilancio sono strumenti di analisi finanziaria utilizzati per valutare la situazione economico-patrimoniale aziendale, indagando sulla capacità di un'impresa di generare reddito negli anni.

Il ROI è un indice che indica il grado di redditività del capitale investito o del ritorno sugli investimenti. Esso è riferito alla gestione caratteristica delle aziende. Si determina dal rapporto tra il reddito operativo (RO) e il totale impieghi. Tale indice evidenzia l'efficienza dei processi della gestione caratteristica e la capacità delle aziende di remunerare il capitale proprio e di terzi. Un ROI più alto del tasso medio di interesse bancario indica un'azienda profittevole a prendere in prestito denaro per investirlo. Viceversa un ROI inferiore ai tassi di interesse bancari mostra un'azienda incapace di generare profitti.

Il ROE è l'indice di redditività del capitale proprio ed esprime il rendimento del capitale conferito dai soci. Tale indice è dato dal rapporto tra il reddito netto e il patrimonio netto. È un indicatore che calcola il grado di economicità ed efficienza dell'impresa.

Il ROS è dato dal rapporto tra il risultato operativo e i ricavi di vendita. Maggiore è il ROS migliore sarà la redditività aziendale in relazione alla capacità di remunerare i flussi dei ricavi. Questo indice è condizionato dai costi aziendali e dai prezzi di mercato e permette di esprime sia l'efficienza interna che quella esterna.

Riferimenti bibliografici

Aluisio S., Specia L., Gasperin C. e Scarton C. 2010, *Readability Assessment for Text Simplification*, Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications: 1-9.

Barbera F., Dagnes J., Salento A. e Spina F. (a cura di) 2016, *Il capitale quotidiano. Un manifesto per l'economia fondamentale*, Donzelli, Roma.

Boumans J.W. e Trilling D. 2016, *Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars*, «Digital Journalism», 4(1): 8–23.

Corrao S. 2005, *L'intervista nella ricerca sociale*, «Quaderni di Sociologia», 38: 147-171.

Di Ciaccio A. e Borra S. 2014, *Statistica. Metodologie per le scienze economiche e sociali*, McGraw-Hill, Milano.

Drisko J. W. e Maschi T. 2016, *Content Analysis*, Oxford University Press, Oxford.

Fabbris L. 2011, *Statistica multivariata*, McGraw-Hill Education Italy, Milano (POD).

Grimmer J. e Stewart B. 2013, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, «Political Analysis», 21(3): 267-297.

Holsti O. 1969, *Content Analysis for the Social Sciences and Humanities*, Addison Wesley, Reading.
Johnson F. e Gupta S.K. 2012, *Web Content Mining Techniques: A Survey*, «International Journal of Computer Applications», 47, 11: 44-50.
Kaufman L. e Rousseeuw, P.J. 2005, *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley & Sons Inc., Hoboken, NJ.
Krippendorff K. 2013, *Content Analysis. An Introduction to Its Methodology*, SAGE Publications, Los Angeles.
Lanzt B. 2015, *Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems*, Packt Publishing, Birmingham.
Liu B. 2010, *Sentiment Analysis and Subjectivity*, in Indurkhya N. e Damerau F.J. (eds.), *Handbook of Natural Language Processing*, CRC Press, Boca Raton: 627-666.
Mayring P. 2000, *Qualitative Content Analysis*, Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 1(2), <http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204>.
Munzert S., Rubba C., Meibner P. e Nyhuis D. 2015, *Automated Data Collection with R. A practical guide to web scraping and text mining*, Wiley, United Kingdom.
Swanson D. R. 1986, *Undiscovered public knowledge*, «Library Quarterly», 56(2): 103-118.
Taboada M., Brooke J., Tofoloski M., Voll K. e Stede M. 2011, *Lexicon-Based Methods for Sentiment Analysis*, «Computational Linguistics», 37, 2: 267-307.
Welbers K., Van Atteveldt W. e Benoit K. 2017, *Text Analysis in R*, «Communication methods and measures», 11, 4: 245-256.
Zani S. e Cerioli A. 2007, *Analisi dei dati e data mining per le decisioni aziendali*, Giuffrè, Milano.